International Journal of Computational Intelligence Systems
Vol. **13(1)**, 2020, *pp.* 1109–1119
DOI: https://doi.org/10.2991/ijcis.d.200728.001; ISSN: 1875-6891; eISSN: 1875-6883
https://www.atlantis-press.com/journals/ijcis/

Research Article

Framework of Computational Intelligence-Enhanced Knowledge Base Construction: Methodology and A Case of Gene-Related Cardiovascular Disease

Yi Zhang[1], Mengjia Wu[1], Hua Lin[2], Steven Tipper[2], Mark Grosser[2], Guangquan Zhang[1], Jie Lu[1,*]

[1]*Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, 61 Broadway, Ultimo, NSW 2007, Australia*
[2]*23Strands, Suite 105, 26 Pirrama Rd, Pyrmont, NSW 2009, Australia*

ARTICLE INFO

Article History
Received 26 Feb 2020
Accepted 03 Jun 2020

Keywords
Bibliometrics
Knowledge management
Computational intelligence
Cardiovascular disease

ABSTRACT

Knowledge base construction (KBC) aims to populate knowledge bases with high-quality information from unstructured data but how to effectively conduct KBC from scientific documents with limited preknowledge is still elusive. This paper proposes a KBC framework by applying computational intelligent techniques through the integration of intelligent bibliometrics—e.g., co-occurrence analysis is used for profiling research topics/domains and identifying key players, and recommending potential collaborators based on the incorporation of a link prediction approach; an approach of scientific evolutionary pathways is exploited to trace the evolution of research topics; and a search engine incorporating with fuzzy logics, word embedding, and genetic algorithm is developed for knowledge searching and ranking. Aiming to examine and demonstrate the reliability of the proposed framework, a case of gene-related cardiovascular diseases is selected, and a knowledge base is constructed, with the validation of domain experts.

1. INTRODUCTION

Knowledge base construction (KBC), known as a process of filling knowledge bases with high-quality information from unstructured data [1], has been a long-standing need in industry sectors, as well as a key research topic in the area of knowledge discovery and management for decades, in which unstructured data is well exploited for information retrieval [2]. In parallel, it is also a common sense that scientific documents such as research articles, patents, and academic proposals contain rich information in science, technology, and innovation, despite potential difficulties in analyzing complicated text data [3]. Given the circumstances, how to conduct KBC from scientific documents becomes attractive for not only researchers but also a wide range of other professions and broad business sectors, which could benefit from a KBC process which leads to improved understanding of emerging domains with limited preknowledge.

Bibliometrics, known as the use of statistical approaches to analyze scientific documents and explore empirical insights for decision support [4], has been used in broad empirical studies for knowledge discovery—e.g., profiling research domains [5], identifying research topics [6], and tracking topic evolution over time [7]. Additionally, aiming to handle issues which result from managing big data such as scalability, uncertainty, and robustness, and domains for which interests on intelligent bibliometrics are raised, emphasizes the refinement required of traditional bibliometric approaches by incorporating intelligent techniques, e.g., topic models, network analytics, neural networks, and other machine learning approaches. Such endeavors include dynamic topic detection and tracking [8], word embedding-incorporated topic extraction [9], streaming data analytics for identifying complicated semantic relationships among research topics over time [10], etc. However, gaps between analytic results of established bibliometric approaches and KBC still exist such as how to design a KBC framework to systematically integrate bibliometric models and effectively manage knowledge, and how to implement the proposed framework for decision support in real-world cases.

Aiming to address the above concerns, this paper proposes a framework of KBC which applies computational intelligence techniques through the integration of intelligent bibliometrics. Oriented to the needs of constructing a knowledge base for emerging research topics—i.e., a domain with insufficient preknowledge in practice but which may contain rich supplementary sources in scientific documents, the authors developed and integrated the following steps: 1) a function of co-word analysis is used to profile a given domain and identify research opportunities through topics that are represented by a set of research terms; 2) a function of coauthorship network and link prediction is developed to recognize key entities (e.g.,

*Corresponding author. Email: jie.lu@uts.edu.au

researchers, institutions, and countries/regions) and recommend potential collaboration partners based on their existing collaborative relationships; 3) a function of scientific evolutionary pathways (SEP) [10] is exploited to trace the evolution of scientific topics, with the use of machine learning techniques and streaming data analytics; and 4) a function of search engine is developed with applicable strategies to involved data sources, incorporating fuzzy logics, word embedding, and genetic algorithm for knowledge representation and ranking.

Aiming to demonstrate the of this KBC framework in a clinically knowledge-rich domain, a case study of gene-related cardiovascular diseases was selected, and a knowledge base was constructed using the functional steps outlined above, which was also validated by reference and iterative feedback from knowledgeable domain experts.

The rest of this paper is organized as follows: Section 2 reviews related work in bibliometrics, KBC, and computational intelligence. Section 3 presents the research framework and related methodologies. A case study of establishing the knowledge base of gene-related cardiovascular diseases is given in Section 4. Technical implications, possible applications, and limitations, as well as future directions, are concluded in Section 5.

2. LITERATURE REVIEW

This section reviews related work from the following two aspects: KBC and bibliometrics.

2.1. Knowledge Base Construction

KBC is defined as the process of populating knowledge bases with high-quality information (e.g., objects, rules, and relationships) from unstructured data [1], and data engineering and machine learning techniques are widely used [11]. Considering KBC as a practical issue, its applications could be traced in broad sectors such as medical science [12], music industry [13], and customer services [14]. There are also a number of KBC systems, acting as a toolkit for general KBC needs—e.g., Elementary [1], TinkerBell [15], and Fonduer [16]. Significantly, Deepdive as a benchmark of KBC provides a database and machine learning-based solution for KBC needs of technology companies, law enforcement agencies, and academic researchers [2]. Similarly but targeting to incomplete knowledge bases, knowledge base augmentation is specifically raised in the area of semantic web, which emphasizes the extraction and identification of entities and relations [17,18].

2.2. Bibliometrics

Pioneered in the early 1960s by Derek Price for observing patterns of scientific activities [19], modern bibliometrics is initially defined as "the application of mathematics and statistical methods to books and other media of communication" [20], and now various data analytic techniques have been incorporated with traditional bibliometric models, involving indicators such as citation/co-citation statistics, word co-occurrence, and coauthorships retrieved from scientific documents [21]. Interactions between knowledge discovery and bibliometrics started decades ago, oriented to specific entities (e.g., research domains, technologies, journals, and regions and countries), in which bibliometric models were applied for transferring raw data to structured knowledge—e.g., identifying topics and relationships [22,23], detecting and tracking emerging trends [10,24], and investigating key players and their collaborative patterns [25,26]. Significantly, machine learning techniques provide new angles and solutions for tasks in knowledge representation, classification, and clustering [9,10,27]. Intelligent bibliometrics are then raised by emphasizing the "development and application of intelligent models for recognizing patterns in bibliometrics" [28].

2.3. Computational Intelligence

Computational intelligence is an area of fundamental research and practical studies exploiting a number of information processing technologies, such as neural networks, fuzzy logics, and evolutionary computation [29]. With the rapid development and wide applications of neural networks, natural language processing (NLP), together with deep learning techniques, creates solutions of comprehensively understanding free text in the real word [30]. Fuzzy logics, rooted in Zadeh's studies in the 1960s [31], helps represent the belongingness of an unknown object to a known category via a grade of membership ranging between 0 and 1. The capability of fuzzy logics in handling uncertainty and fuzzy features have been well recognized [32]—e.g., representing customer comments [33] and describing user preferences [34] in the applications of recommender systems. Additionally, fuzzy linguistics emphasizes the use of fuzzy logics in transferring subjective semantics into numeric numbers and have been widely applied in a broad range of industry and system management practices [35]. Evolutionary computation closely relates to optimization problems, which provides further flexibility, adaptability, and robustness to problem-solving [36]. Our paper, oriented to the task of KBC, incorporates techniques of computational intelligence in supporting bibliometric models—e.g., word embedding techniques (with shallow deep learning techniques) which are introduced for knowledge representation, the membership grades of fuzzy sets are used as an indicator for ranking, and a genetic algorithm is exploited in handling a task of multi-objective ranking.

3. METHODOLOGY

This paper aims to propose a framework of base construction (KBC) for scientific and technological domains by applying computational intelligence techniques through integrating intelligent bibliometrics. The designed framework is given in Figure 1. which is oriented to scientific documents, so as a result our framework includes data collection and preprocessing and KBC that contains four functions: 1) a function of topic analysis is designed for profiling knowledge landscapes via research topics and their relationships; 2) a function of network analytics is exploited to identify key players in a given knowledge domain and detect potential collaborations among those key players through link prediction; 3) a function of SEP involving streaming data analytics and machine learning techniques is proposed for tracking knowledge trends and predicting emergent research topics; and 4) a function of knowledge searching and ranking is developed to filter scientific documents from raw datasets, based on given criteria from domain experts, combined with fuzzy logics, word embedding, and optimization

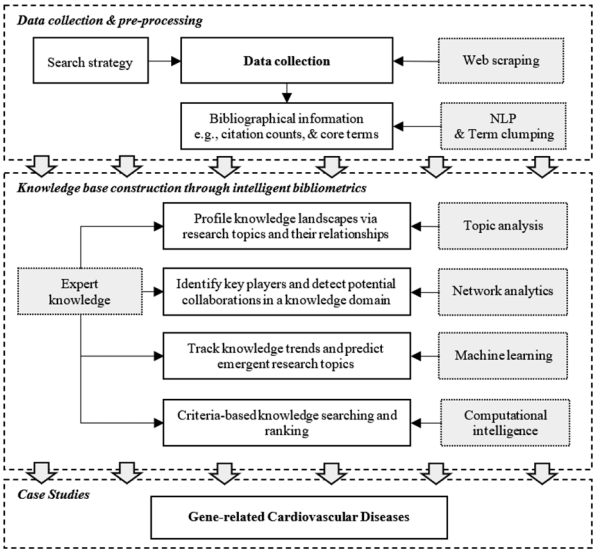

Figure 1 | Framework of computational intelligence-enhanced knowledge base construction.

techniques used to assist in knowledge representation and similarity measurements in the final analysis.

3.1. Data Collection and Preprocessing

This study focused on a number of data sources of scientific documents which we could investigate and may benefit from future research in using this framework, such as the Web of Science (WoS) database and the PubMed database for academic articles, Derwent World Patent Index (DWPI) database and the United States Patent and Trademark Office (USPTO) database for patents, and the National Science Foundation of the United States (NSF-US) database for academic proposals. From a methodology perspective, targeting specific knowledge domains in science and technology required a search strategy to collect relevant scientific documents and then the bibliographical information of scientific documents, such as titles, abstracts, keywords, authors, and their affiliations, will be extracted and analyzed. Specifically, NLP techniques can be applied to retrieve terms (including words and phrases) from the free text of titles and abstracts, and a term clumping processing [37] is involved in removing noises and consolidating synonyms.

The outputs of this phase include lists of authors, affiliations, countries/regions, and cleaned terms, and each list also contain the number of scientific documents associated with these items.

3.2. Knowledge Base Construction

The KBC process is designed by integrating four specific functions. With aid of co-occurrence analysis, text segmentation is used to retrieve knowledge by extracting research topics and authorship is exploited to identify key players and their collaboration patterns. An approach of SEP is refined and adapted to tracking knowledge trends and predicting emergent topics through machine learning and streaming data analytics. Then, aiming to represent retrieved knowledge, computational intelligence techniques including word embedding, fuzzy sets, and genetic algorithms are involved in knowledge searching and ranking.

3.2.1. Profiling knowledge landscapes

While a topic is defined a collection of terms representing similar semantic meanings, the knowledge landscape of a given domain is described as a set of topics and their relationships. Targeting to the list of cleaned terms, co-occurrence analysis is applied for initially measuring the relationships between terms [38], with a hypothesis that if two words frequently appear together, they are similar [39]. The corresponding algorithm is described below:

- Given that $T = \{t_1, \ldots, t_i, \ldots, t_n\}$ is the list of cleaned terms and term t_i could represented by a vector $C_1 = \{c_{1,i}, \ldots, c_{i,i} \ldots, c_{i,j}, \ldots, c_{n,i}\}$, in which $c_{i,j}$ represents the frequency of the co-occurrence between terms t_i and t_j in the dataset;
- The similarity $Sim\left(t_i, t_j\right)$ between terms t_i and t_j could be calculated by the Salton's cosine [40], i.e.,

$$Sim\left(t_i, t_j\right) = \frac{C_i \cdot C_j}{|C_i|\,|C_j|}$$

- A triangle similarity matrix S_T is then generated, which records the similarities between all pairs of terms in list T;
- A network $G_t\left(T, E_t\right)$ is constructed based on matrix S_T, in which each term is represented by a node while E_t represents the set of edges in the network and is filled with the similarities $Sim\left(t_i, t_j\right)$.
- Regarding to the map of science [41], an approach of community detection is applied to visualize the network $G_t\left(T, E_t\right)$ through certain communities, which are then identified as topics in a given domain.

Note that despite numerous text similarity algorithms in the literature, our pilot studies have examined that the cosine similarity algorithm could achieve the best performance in bibliometric datasets [42], and thus, the entire study of this paper exploits Salton's cosine algorithm for similarity measurements.

The science map as well as the network $G_t\left(T, E_t\right)$ are considered as the outcomes of this function, which profile knowledge landscapes of a given domain in a vivid manner. Additionally, the structure of communities and terms provides a hierarchical solution to represent and organize knowledge and helps understand a domain at a macro level.

3.2.2. Identifying key players and detecting potential collaborations

The identification of key players follows the algorithm of co-occurrence analysis presented above but using lists of cleaned entities such as authors A, research institutions O, and countries/regions R. Thus, the generated networks are denoted as $G_a\left(A, E_a\right)$, $G_o\left(O, E_o\right)$ and $G_r\left(R, E_r\right)$ respectively. Such science maps and networks visualize key players and their existing collaborations, as well as their research groups that are identified as communities in the maps.

A link prediction approach [43] is applied for analyzing the topological structure of the generated networks to detect potential collaborations for authors and research institutions. The basic assumption

of link prediction approaches is that: if A and B links with C respectively, A will link with B in the near future, and thus the task of link prediction approaches is to fulfill those missing links with weights.

Considering two types of collaborative strategies—i.e., maintaining existing collaborations and establishing new collaborations, two lists of recommendation pairs will be provided, in which one is for existing collaborators and the other is for potential collaborators with whom they have never collaborated before.

3.2.3. Tracking knowledge trends and predicting emergent topics

A general concern for knowledge trends is the potential change underlying certain given knowledge over time. For example, in the early 2000s and earlier, "data mining" mostly referred to studies in database management and data warehouse, but now "data mining" closely interacts with machine learning techniques which were not common in those early days. Given this evolutionary history occurs, our functional aim is to track knowledge trends and predict emergent topics by considering such changes. The main function is refined from the approach of SEP [10], for which related definitions are given below:

Definition 1. A topic[1] $Tp\,(c, r, S)$ is a collection of scientific documents (we use "record" in the following description) and is geometrically represented as a circle, in which c is its centroid identified as the mean of the vectors of all involved articles a, r is its boundary identified as the largest Euclidean distance between c and all other articles, and S indicates its time slice.

Definition 2. Regarding a concept of "sleeping beauties" [44] for identifying scientific innovation from the literature, a topic could be either alive or dead.

Definition 3. The data corpus Φ consists of k time slices S, and each time slice includes n articles.

The stepwise algorithm is described as follows:

Step 1: Set S_0 as the initial time slice, and group all involved articles as one topic $Tp\,(c, r, S_0)$ and consider it as the initial topic of an evolutionary pathway.

Step 2: Iteratively process the data corpus Φ as a simulated stream—i.e., process one time slice once by analyzing its articles one by one.

Step 3: Measure the similarity $sim\,(a, Tp)$ between a forthcoming article a and centroids of all alive topics via Salton's cosine [40]—i.e.,

$$Sim\,(a, Tp) = \frac{a \cdot c}{|a||c|}$$

Step 4: Assign article a to the most similar topic Tp, and calculate the Euclidean distance $E\,(a, c)$ between a and the topic's centroid c. If $E\,(a, c) < r$, the article will be directly assigned to the topic, or else, the article will be labeled as "drift." Then, return to Step 3 and analyze the next article.

Step 5: At the end of each iteration, set a topic as "dead" if it does not receive any articles in two continuous time slices, which means since this topic is generated, there is not any new knowledge accumulated to it in the following time slice. Then, logically "live" topics continue, given that it is in time slice S_x, we iteratively apply an unsupervised K-means approach [3] to each topic and group their assigned articles labeled with "drift" into certain sub-topics $Tp'\,(c', r', S_x)$.

Step 6: Measure the cosine similarity between Tp' and two sets of topics—the assigned topic Tp and all dead topics Tp_d. If $Sim\,(Tp', Tp) > Sim\,(Tp', Tp_d)$, the relationship between Tp and Tp' is defined as "descendent-predecessor," or else, the "dead" topic Tp_d will be resurged and set as "alive," and then, becomes the predecessor of topic Tp'.

Step 7: Label topic Tp' via the term with the highest similarities with all other terms in this topic—if the term has already been used by existing topics, choose its following terms.

Step 8: Update all alive topics by recalculating their centroid and boundary and return to Step 2 until the stream ends.

The outcome of the SEP approach is a list of topics with information such as labels, descriptions, involved terms and articles, and "sleeping beauties"-related indicators (e.g., born time, dead time, and resurgence). These topics could be visualized in a direct network, in which each topic is represented by a node and weighted edges represent the similarities between their connected nodes. It is clear that this network provides a solution of tracing knowledge trends by detecting such predecessor–descendant relationships and predicting emerging topics by identifying topics of "sleeping beauties."

Note that the SEP approach exploited in this study mostly follows the version presented in [10] but aiming to further omit human intervention and adapt to practical needs by consulting with certain domain experts, we modified the algorithm from the following aspects: 1) in Step 1, only one initial topic is grouped rather than applying a K-means approach to group records in a given number of topics; and 2) an unsupervised K-means approach is applied to take the place of the hierarchical clustering approach in Step 5.

3.2.4. Criteria-based knowledge searching and ranking

The function of criteria-based knowledge searching and ranking is designed to conduct data argumentation with limited expert knowledge. The input of the function is a core collection of scientific documents, which may also be manually collected by domain experts and indicates their existing knowledge base,[2] and the task of this function is to extensively collect relevant articles from the entire dataset based on the core collection and return a set of relevant articles.

Two ranking lists are initially generated—one is based on the cosine similarities between the vectors of individual articles and the core collection, which are created by a Doc2Vec model [45]; and the

[1]Note that despite the fact that Sections 3.1.1 and 3.1.3 use the term "topic," but in Section 3.1.1 a topic is a set of terms while Section 3.1.3 highlights that a topic is a set of scientific documents. There is not any conflict between the two definitions which are two types of definitions for topics in bibliometrics.

[2]This setting is motivated by the finding that clinical practices usually record a relatively small number of academic articles collected for their on-going cases. Thus, this function is to automatically extend this small group of articles into a well-established dataset of similar articles.

other is based on the grade of membership that an individual article belongs to the core collection, in which fuzzy sets are involved. Then, the nondominated sorting genetic algorithm II (NSGA-II) [46] is exploited to identify *nonnominated solutions* (i.e., considering a multi-objective optimization problem, solutions that are superior to the rest of solutions when all objectives are considered, but are inferior to other solutions in one or several objectives [47]), considering both ranking criteria. The stepwise algorithm of this function is described as follows:

Step 1: A Doc2Vec model based on the Word2Vec model [45] is applied to the entire data corpus (including the core collection Δ provided by domain experts), and each article a is represented by an abstract vector.

Step 2: A search strategy consisting of a set of search terms is provided by users, and a combinative search is conducted to return a set P of articles a' that coincide with the search strategy.

Step 3-1: The mean M_c of the core collection is calculated as $M_c = \sum a'$, and then measure the cosine similarity between M_c and a'. Return a ranking list R_1 based on the similarities.

Step 3-2: Given that $\Psi_{a'}$ and Ψ_Δ is the set of specific tags (e.g., Mesh terms and keywords) of one searched article and the core collection respectively, and $F_\Delta(a')$ is the membership grade that article a' belongs to the core collection. The membership function F_Δ is defined as follows. Then, rank articles a' based on the membership grades and get a ranking list R_2.

$$F_\Delta(a') = \frac{|\Psi_{a'} \cap \Psi_\Delta|}{|\Psi_\Delta|}$$

Step 4: Exploit the fast-nondominated-sort approach of NSGA-II [46] to calculate the number of articles $d_{a'}$ that each article dominates (i.e., inferior to article a' in both ranking lists), and then rank all the searched articles based on $d_{a'}$—i.e., the larger the higher, the integrated calculation procedure is given in Algorithm 1.

Algorithm 1: Fast-nondominated-sort approach of NSGA-II to rank the record results

Input: Set P of returned articles, Ranking list R_1, and Ranking list R_2
Output: Comprehensive Ranking R_3
Steps:
1 For article a' in P:
2 if $R_2[a'] != 0$: ← where $R_2[a']$ is the ranking value of a' in R_2
3 $R_3[a'] =$ The number of mutual articles which are ranked behind a' both in R_1 and R_2
4 else:
5 $R_3[a'] =$ The num of articles ranked behind a' in R_1) / 2
6 $R_3 =$ P ranked by $R_3[a']$ for article a' in P
7 end

The function provides a solution of augmenting a specified knowledge base with limited expert support, in which semantic similarities between scientific articles are exploited.

4. CASE STUDY

Aiming to demonstrate the reliability of the proposed framework for KBC, a knowledge base for gene-related cardiovascular diseases was constructed.

4.1. Search Strategy and Data Preprocessing

Cardiovascular disease has become a key concern in the modern world and discovering the relationships between cardiovascular diseases and human genes would be a key to provide new angles to potentially diagnose, curatively treat or manage such diseases. The explosion of literature in this field is evident via a simple Google search of the public web domain using the simple search ["cardiovascular disease" AND "genetics"] which returns "About 24,600,000 results (0.43 seconds)" Thus, gene-related cardiovascular disease has been an emerging topic in medical science, genomics, and related disciplines [48].

The PubMed database[3] owned by the US National Library of Medicine and National Institutes of Health is an open-access database that includes more than 30 million items of biomedical literature from Medline, life science journals and online books. It has been widely used as a search engine and data source for both academic research and professional practice sectors, and thus we decided to choose the PubMed as the target database.

When considering search strategies, accuracy and coverage are the two key foci of proposing empirical search strategies, so this case study set coverage as the priority, and the proposed search strategy that combines MeSH (Medical Subject Headings) terms, which are specific tags in PubMed database, and free text in combinations as follows:

(``Cardiovascular Diseases/genetics" [Mesh] OR ``Cardiovascular Diseases" [Mesh] OR CVD OR Cardiovascular* disease*) AND (``Genetic Phenomena" [Mesh] OR Genome* OR Gene OR Genetic* OR DNA OR RNA) AND (``2008/01/01" [PDat]: ``3000/12/31" [PDat])

With the aid of VantagePoint,[4] the data-preprocessing was conducted from two aspects: the removal and consolidation of terms (including words and phrases), and the disambiguation of author/affiliation names.

An NLP function[5] integrated with the VantagePoint was applied to the combined fields of titles and abstracts of the 142,877 articles, and retrieved 2,340,100 terms. A term clumping process step

[3]More information can be found on the website: https://www.ncbi.nlm.nih.gov/pubmed/

[4]VantagePoint is commercial software used in text mining and particularly in science, technology, and innovation text analysis. Its involvement in this study includes its NLP function and a light AND function. More details can be found on the website: https://www.thevantagepoint.com/

[5]Note that since NLP is not a key focus of this study, in this paper we simply used the NLP function integrated in VantagePoint. However, note that any existing NLP functions could adapt to the proposed methods.

was then used to remove noisy terms and consolidate technical synonyms, from which 264,125 terms were identified as the key terms of representing the knowledge base of gene-related cardiovascular diseases. The stepwise results of the term clumping processing is given in Table 1.

An author name disambiguation (AND) function was applied to clean authors and affiliations. Its main foci include: 1) to consolidate different presenting formats of the same authors—e.g., "Boerwinkle, Eric" and "Eric Boerwinkle"; 2) to consolidate the names of all branches into the name of their head quarter—e.g., "St Vincent's Hospital, Sydney" and "St Vincent's Hospital, Melbourne"; and 3) to remove authors and affiliations that only appear once in the dataset, since they will not have any co-occurrence instance. Eventually, 162,817 authors and 11,321 affiliations were retrieved from their raw lists, with 496,178 author names and 327,216 affiliation names respectively. In particular, 117 Australian affiliations were identified from a raw 607-item list. The lists of cleaned authors and affiliations would be the main inputs of identifying key players in gene-related cardiovascular diseases.

4.2. KBC for Gene-Related Cardiovascular Diseases

The construction of the knowledge base for gene-related cardiovascular diseases includes profiling knowledge landscapes, identifying key players and detecting potential collaborations, tracking knowledge trends and predicting emergent topics, and criteria-based knowledge searching and ranking.

4.2.1. Profiling knowledge landscapes

A co-term network $G_t(T, E_t)$ was constructed, including the top 5000 high frequency terms and 1,735,189 edges, with the aid of VoSViewer [41],[6] a smart local moving algorithm was applied for visualizing the network, as given in Figure 2.

According to Figure 2, the knowledge landscapes of gene-related cardiovascular diseases are illustrated by four core groups: coronary artery diseases (blue nodes), atrial fibrillation (green nodes), molecular mechanisms and heart failure (yellow nodes), and protective effects (red nodes). If targeting one specific node (e.g., atrial fibrillation), its co-occurrent relationships with other nodes could be observed as well, as given in Figure 3.

It is clear that this function provides a vivid way to profile the knowledge landscapes of a given domain by identifying core topics and their relationships.

4.2.2. Identifying key players and detecting potential collaborations

Three coauthorship networks $G_a(A, E_a)$, $G_o(O, E_o)$ and $G_r(R, E_r)$ were generated respectively for individual researchers, research institutions, and countries/regions of gene-related cardiovascular diseases. As an example, the country-based coauthorship network $G_r(R, E_r)$ is given in Figure 4.

As shown in Figure 4, the United States, the United Kingdoms, France, the Netherlands, Australia, and China are leading the research of gene-related cardiovascular diseases, and the strengths

Table 1 Stepwise results of the term clumping process.

Step	Description	#T
0	Raw terms retrieved through NLP	2,340,100
1	Remove single-word terms, e.g., "information"	2,121,328
2	Remove terms starting/ending with nonalphabetic characters, e.g., "step 1" and "1.5 m/s"	2,106,207
3	Remove meaningless terms, e.g., pronouns, prepositions, and conjunctions	2,104,160
4	Remove common terms in scientific articles, e.g., "research framework"	2,080,349
5	Remove terms appearing in only one record	317,105
6	Consolidate terms with the same stem, e.g., "information system" and "information systems"	264,261
7	Consolidate synonyms based on expert knowledge, e.g., "co-word analysis" and "word co-occurrence analysis"	234,719
8	Consolidate synonyms based on given rules, e.g., removing terms starting with "existing"	201,512
9	Remove terms appearing less than 4 times;	81,581

NLP, natural language processing.
Note that #T: the number of terms.

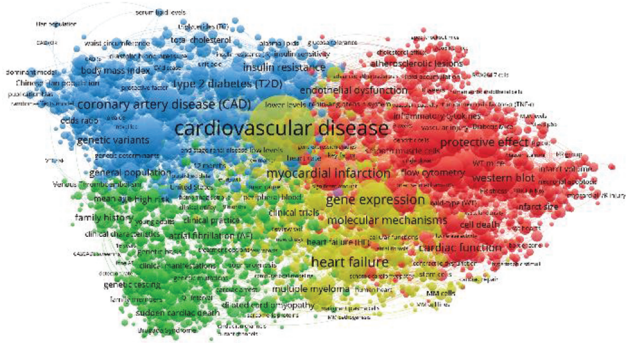

Figure 2 Co-term network for profiling the knowledge landscape of gene-related cardiovascular diseases.

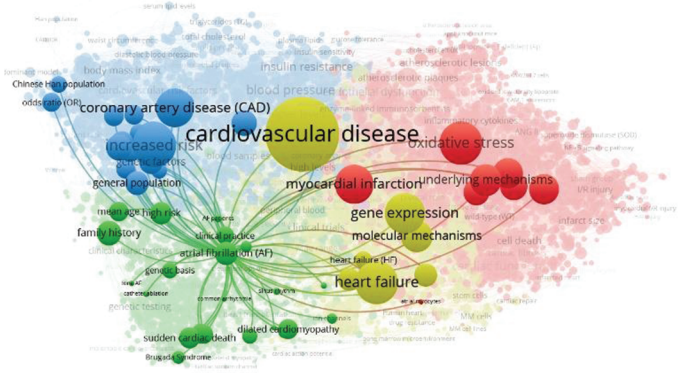

Figure 3 Co-occurrence relationships of atrial fibrillation.

[6]The default setting of parameters in VoSViewer was exploited, which has been examined in the above cited reference.

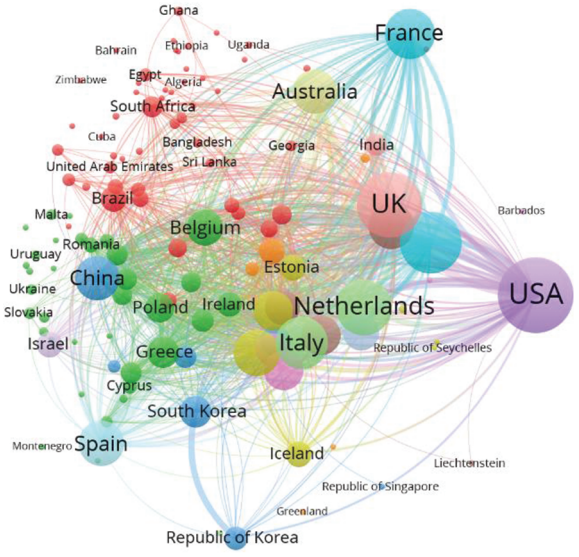

Figure 4 | Country-based coauthorship network for identifying key players in gene-related cardiovascular diseases.

of their collaborations with other countries could be observed as well. Similarly, based on the coauthorships, active researchers, institutions, and countries with strong collaborations with others could be identified as the key players in a given domain.

Furthermore, since a coauthorship network could meet with the requirements of complex networks [49], such as scale free and small world, network analytics could also be applied to further analyze the topological structure of the network. Considering an institution-based coauthorship network $G_o(O, E_o)$ as an example and choosing St Vincent's Hospital (Australia)[7] as a target institution, a resource allocation approach (as is has been proved that this approach achieves better performance than other algorithms like Adamic/Adar index and common neighbors [50]) was applied to fulfill missing links in $G_o(O, E_o)$, which then could be facilitated to detect potential collaborations between unconnected nodes. The predicted further collaborators and potential collaborators in the area of gene-related cardiovascular diseases for St. Vincent's Hospital are given in Table 2.

This function creates a way to investigate the key players of a given domain by identifying who they are, exploring how they collaborate with each other, and predicting how such collaborations will be in the near future.

4.2.3. Tracking knowledge trends and predicting emergent topics

The refined SEP approach was applied to the 142,877 articles, which were simulated into a bibliometric data stream with 12 time slices, covering the time period from 2008 to 2019. The SEP for tracking knowledge trends of gene-related cardiovascular diseases is given in Figure 5.

[7]St Vincent's Hospital has several independent branches in Australia, such as St Vincent Melbourne, and St Vincent Sydney, but in this study, we combined all these branches as "St Vincent's Hospital."

Table 2 | Top 5 predicted further collaborators who have collaborated before and top 5 predicted new collaborators in gene-related cardiovascular diseases for St Vincent's Hospital in the world.

	Further Collaborators	Pre. Str.	Est. Coll.
1	Harvard Medical School	0.94	9
2	University College London	0.90	3
3	University of Oxford	0.89	8
4	Massachusetts General Hospital	0.86	5
5	Brigham and Women's Hospital	0.85	7
	New Collaborators	**Pre. Str.**	**Est. Coll.**
1	University of Michigan	0.61	0
2	Washington University School of Medicine	0.57	0
3	Radboud University Medical Center	0.55	0
4	Queen Mary University of London	0.54	0
5	Maastricht University	0.52	0

Note that Pre. Str.: Predicted Collaborative Strength; and Est. Coll.: Existing Collaborations.

208 nodes and 207 edges—i.e., their descendent-predecessor relationships were generated. Each node represents a specific topic, with its detailed information on the year of born, the number of articles, the number of terms, the value of term frequency inverse document frequency (TFIDF) analysis, and the number of survival batches. Specifically, regarding the design of "sleeping beauty" detection, four types of nodes were classified:

- *Always alive*, indicating a continuous interest from the community to this topic.
- *Resurgence and alive*—i.e., the "sleeping beauties" which became dead topics, then were awaken by the involvement of possible new concepts, techniques, materials, and devices, and are still a core interest of the community.
- *Dead with resurgence*, indicating former "sleeping beauties" but have already become dead topics again.
- *Dead without resurgence*, indicating topics either which are meaningless to the community or whose potential has not been discovered yet.

The descriptive statistics of nodes in the four types are given in Table 3. As shown, topics in *resurgence and alive* have the highest average value of TFIDF, indicating its relatively high-quality and emerging potential.

For a reference, the 14 topics in *resurgence and alive* and their related information are provided in Table 4, and coinciding with the concept of "sleeping beauties," these topics are considered as emergent topics in the area of gene-related cardiovascular diseases.

This function exploits streaming data analytics and machine learning techniques to draw the knowledge trends of a given domain and the involvement of "sleeping beauty" detection creates a manner to predict emergent research topics in the area.

Aiming to compare the benefits of the proposed method, we list 20 terms with the highest term frequency in Table 5. These 20 terms were collected from the term clumping process (i.e., Step 9 in Table 1) and were normally considered as the outcome of traditional bibliometrics for identifying key research topics.

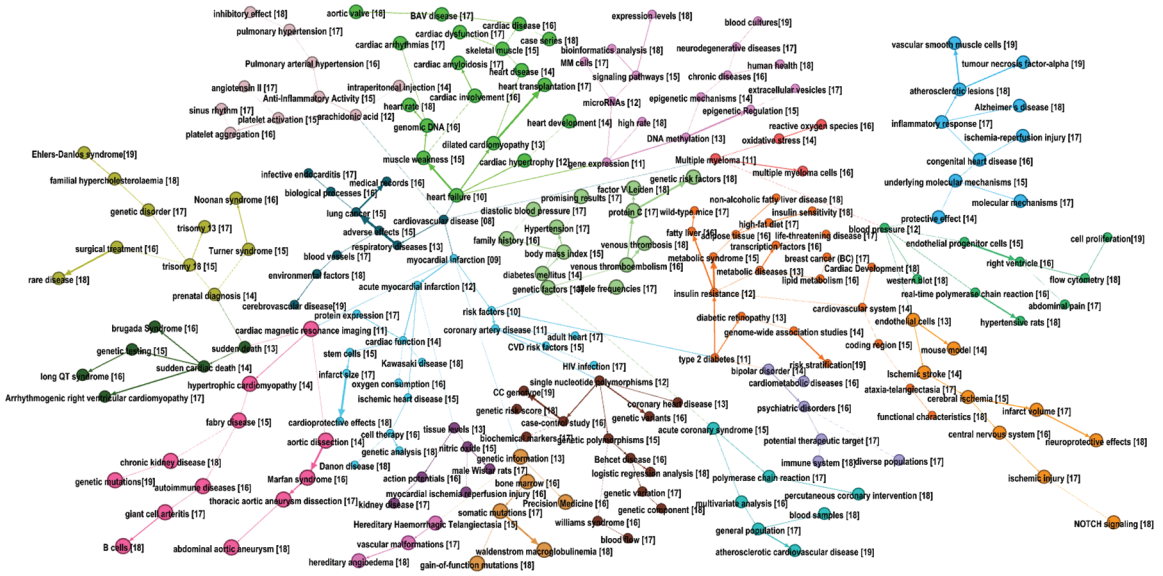

Figure 5 | Scientific evolutionary pathways for tracking knowledge trends of gene-related cardiovascular diseases.

Table 3 | Descriptive statistics of SEP topics in four types.

Types	#N	Max	Min	Avg.	S.D
Always alive	148	1.396	0.028	0.476	0.374
Res. and alive	14	1.400	0.203	0.766	0.405
Dead with res.	3	0.450	0.142	0.334	0.167
Dead without res.	43	0.890	0.017	0.276	0.219

SEP, scientific evolutionary pathways.
Note that #N: number of nodes; Max, Min, Avg. and S.D: the maximum, minimum, average, and standard deviation of the term frequency inverse document frequency (TFIDF) values.

Table 4 | "Sleeping beauty" detection for predicting emergent topics in gene-related cardiovascular diseases.

Category	Born	TFIDF	#S
Risk factors	2010	1.244	9
Coronary artery disease	2011	1.229	7
Single nucleotide polymorphisms	2012	1.024	6
Sudden death	2013	0.570	4
Endothelial cells	2013	1.275	6
Metabolic diseases	2013	0.818	5
Prenatal diagnosis	2014	0.696	3
Protective effect	2014	1.400	3
Genome-wide association studies	2014	0.495	3
Stem cells	2015	0.707	4
Biological processes	2016	0.442	3
Multiple myeloma cells	2016	0.279	3
Action potentials	2016	0.337	3
Cardiac involvement	2016	0.203	3

TFIDF, term frequency inverse document frequency.
Note that #S: the number of survival years.

It is clear that the key outcome of the proposed method is Figure 5, in which all terms in Table 5 could be identified, and then Table 4 provides a supplementary source to identify emerging topics (i.e., significant and within increasing interests to the community). Initially, the overlaps between Tables 4 and 5 (e.g., "risk factors" and "coronary artery disease") could the potential cross-boundaries in emerging topics and key topics in the area of gene-related cardiovascular diseases. We conducted an extensive consultation with our

Table 5 | Top 20 terms with the highest term frequency on gene-related cardiovascular diseases.

No.	Term	No.	Term
1	Cardiovascular disease	11	Endothelial cells
2	Heart failure	12	Multiple myeloma
3	Myocardial infarction	13	Protective effect
4	Oxidative stress	14	Ischemic stroke
5	Gene expression	15	Molecular mechanisms
6	Risk factors	16	Underlying mechanisms
7	Coronary artery disease	17	Cardiac function
8	Type 2 diabetes	18	Diabetes mellitus
9	Blood pressure	19	Mouse model
10	Cardiovascular disease risk factors	20	Protein expression

clinical and research experts, and they agreed with the results based on their expertise and domain knowledge. They also concluded the following observations and may anticipate our further studies: 1) the life of "sleeping beauty" terms may be due to more specific knowledge requiring changes in use of terminology to define the areas of emergent interest in clinical and research areas; 2) they expressed no surprise that these topics and terms are evolving and appear interesting for further research in papers such as this and for validation in clinical settings i.e., the findings from the use of the KBC are not discordant with clinical experts understanding of the topics.

4.2.4. Criteria-based knowledge searching and ranking

A user-friend interface was specifically designed for this function, which consists of 2 sets of auto-completed input textboxes for search terms, an input textbox for the expert knowledge-based core collection, and an output panel for listing ranked relevant documents searched from the entire dataset. The interface is given in Figure 6. Specifically, there are five steps to utilize this interface:

Step 1: Import the full knowledge base—the dataset;

Step 2: Select a group of entities (including diseases, symptoms, drugs, etc.), terms and/or another group of gene terms as inputs,

Figure 6 Interface for the function of criteria-based knowledge searching and ranking.

those selected terms are automatically combined by "OR" inside the group and "AND" between the group to form search criteria, representing the co-occurrence of 2 term-groups from a bibliometric perspective.

Step 3: Click button "search," articles which meet with the search criteria will be retrieved and related bibliographic data is displayed on the bottom panel.

Step 4: Pick out a set of core collection which contains one or more article(s) from the dataset based on expert opinion and type in their PubMed IDs, then click button "Confirm," the details of picked set will be displayed.

Step 5: Click button "Calculate Semantic Similarity" to calculate the similarity of every single article in search results with the set of core collection based on a pretrained doc2vec model. All the searching results will be redisplayed and ranked by the similarity from high to low in the text field of "Semantic Ranking Results."

Step 6: Click button "Calculate MeSH Similarity" to calculate the similarity of every single article in search results with the set of core collection based on the num of mutual their MeSH terms. All the searching results will be redisplayed and ranked by the num of mutual MeSH terms from high to low in the text field of "MeSH Ranking Results."

Step 7: Click button "Comprehensive Ranking" to apply fast-nondominated-sort approach of NSGA-II to the semantic and MeSH ranking lists, a new comprehensive ranking list would be generated. All the searching results will be redisplayed and ranked by the new ranking value from high to low in the text field of "Comprehensive Ranking Results."

This function directly interacts with users and provides a solution of knowledge search and augmenting an existing knowledge base with limited expert support.

5. DISCUSSION AND CONCLUSIONS

This paper proposes a methodology for developing a KBC framework by applying computational intelligent techniques through the integration of intelligent bibliometrics. Specifically, co-word and coauthorship statistics were exploited for profiling research domains, and identifying research topics and key players; network analytics (e.g., link prediction approaches) were integrated for analyzing the topological structures of generated networks—e.g., recommending potential collaborators for individual researchers and research institutions; streaming data analytics and learning techniques were integrated for tracking knowledge trends and predicting emergent topics; and the word2vec model was facilitated for measuring the similarities between scientific articles. The demonstration of this framework in the case of gene-related cardiovascular diseases illustrates the feasibility and reliability of the proposed method. Despite that certain existing techniques in bibliometrics and computational intelligence were exploited in the proposed framework, modifications were conducted to adapt to practical needs, such as the reduction of human intervention and computational cost.

It is clear that KBC, as well as knowledge base argumentation, is a fundamental step for developing recommender systems and other practical applications. That is to say, this framework provides such a solution of constructing and augmenting knowledge bases with limited expert support and in an express manner, which could be helpful in a wide range of application domains where knowledge management and information systems can be usefully applied to solve big data analysis problems. Additionally, despite a specific focus on the areas of science and technology, with the aid of external data sources, such as social media (e.g., Twitter), the framework could be adapted to relatively general cases, such as user preference-based behavior analysis and policy-oriented sentiment analysis.

Certain limitations of the proposed framework are noted and remain for future studies, including 1) this framework concentrates on the use of bibliographical information of scientific articles based on abstracts and titles which, despite being valuable and commonly acceptable for initial or quick review of articles, may not represent the possible complications in analyzing full-text articles where the analysis of the entire article may create a bonus for uncovering sentiments and semantics; 2) the construction of a new knowledge base inherently lacks proven validation measurements, and thus future research should develop approaches that combine qualitative and quantitative methodologies for evaluating KBC methodology and models of performance from multiple aspects, such as knowledge coverage, the rate of information missing, and the accuracy of topic extraction and further relationship identification; and 3) since this is still a framework, further investigation is required to connect it with studies of recommender systems and other applications.

ACKNOWLEDGMENTS

This work is supported by the industry project between the University of Technology Sydney and the 23 Strands Pty Ltd (#PRO19-8923) and the

Australian Research Council under Discovery Early Career Researcher Award DE190100994.

CONFLICT OF INTEREST

There is not any conflict of interest in this work.

AUTHOR CONTRIBUTIONS

Designed research: Yi Zhang, Hua Lin, Mark Grosser, Guangquan Zhang, Jie Lu, Mengjia Wu Data collection: Yi Zhang, Mengjia Wu, Hua Lin, Mark Grosser Data analysis: Yi Zhang, Mengjia Wu, Hua Lin Data visualization and toolkit development: Mengjia Wu, Yi Zhang Results interpretation and validation: Steven Tipper, Yi Zhang Wrote the paper: Yi Zhang, Mengjia Wu

REFERENCES

[1] F. Niu, C. Zhang, C. Ré, J. Shavlik, Elementary: large-scale knowledge-base construction via machine learning and statistical inference, Int. J. Semant. Web Inf. Syst. 8 (2012), 42–73.
[2] C. De Sa, *et al.*, Deepdive: declarative knowledge base construction, ACM SIGMOD Record. 45 (2016), 60–67.
[3] Y. Zhang, G. Zhang, H. Chen, A.L. Porter, D. Zhu, J. Lu, Topic analysis and forecasting for science, technology and innovation: methodology and a case study focusing on big data research, Technol. Forecast. Soc. Change. 105 (2016), 179–191.
[4] W. Hood, C. Wilson, The literature of bibliometrics, scientometrics, and informetrics, Scientometrics. 52 (2001), 291–314.
[5] Y. Guo, L. Huang, A.L. Porter, Profiling research patterns for a new and emerging science and technology: dye-sensitized solar cells, in 2009 Atlanta Conference on Science and Innovation Policy, Atlanta, GA, USA, 2009, pp. 1–7.
[6] C.-K. Yau, A. Porter, N. Newman, A. Suominen, Clustering scientific documents with topic modeling, Scientometrics. 100 (2014), 767–786.
[7] Y. Huang, *et al.*, A hybrid method to trace technology evolution pathways: a case study of 3D printing, Scientometrics. 111 (2017), 185–204.
[8] W. Ding, C. Chen, Dynamic topic detection and tracking: a comparison of HDP, C-word, and cocitation methods, J. Assoc. Inf. Sci. Technol. 65 (2014), 2084–2097.
[9] Y. Zhang, *et al.*, Does deep learning help topic extraction? A kernel k-means clustering method with word embedding, J. Informet. 12 (2018), 1099–1117.
[10] Y. Zhang, G. Zhang, D. Zhu, J. Lu, Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics, J. Assoc. Inf. Sci. Technol. 68 (2017), 1925–1939.
[11] C. Zhang, *et al.*, DeepDive: declarative knowledge base construction, Commun. ACM. 60 (2017), 93–102.
[12] A.B. McCoy, *et al.*, Development and evaluation of a crowdsourcing methodology for knowledge base construction: identifying relationships between clinical problems and medications, J. Am. Med. Informat. Assoc. 19 (2012), 713–718.
[13] S. Oramas, L. Espinosa-Anke, M. Sordo, H. Saggion, X. Serra, Information extraction for knowledge base construction in the music domain, Data Knowl. Eng. 106 (2016), 70–83.
[14] B. Pereira, C. Robin, T. Daudert, J.P. McCrae, P. Mohanty, P. Buitelaar, Taxonomy extraction for customer service knowledge base construction, in International Conference on Semantic Systems, Karlsruhe, Germany, 2019, pp. 175–190.
[15] M. Al-Badrashiny *et al.,* TinkerBell: Cross-lingual cold-start Q45 knowledge base construction, in Text Analysis Conference, Gaithersburg, Maryland, USA, 2017.
[16] S. Wu, *et al.*, Fonduer: knowledge base construction from richly formatted data, in Proceedings of the 2018 International Conference on Management of Data, ACM, Houston, TX, USA, 2018, pp. 1301–1316.
[17] D. Ritze, O. Lehmberg, Y. Oulabi, C. Bizer, Profiling the potential of web tables for augmenting cross-domain knowledge bases, in Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, Montréal, Canada, 2016, pp. 251–261.
[18] R. Yu, U. Gadiraju, B. Fetahu, O. Lehmberg, D. Ritze, S. Dietze, KnowMore–knowledge base augmentation with structured web markup, Semantic Web. 10 (2019), 159–180.
[19] D. Price, Little Science, Big Science, Columbia University Press, New York, NY, USA, 1963.
[20] A. Pritchard, Statistical bibliography or bibliometrics, J. Document. 25 (1969), 348–349.
[21] Y. Zhang, Y. Guo, X. Wang, D. Zhu, A.L. Porter, A hybrid visualisation model for technology roadmapping: bibliometrics, qualitative methodology and empirical study, Technol. Anal. Strat. Manag. 25 (2013), 707–724.
[22] R.M. Shiffrin, K. Börner, Mapping knowledge domains, Proc. Natl. Acad. Sci. 101 (2004), 5183–5185.
[23] Y. Zhang, H. Chen, J. Lu, G. Zhang, Detecting and predicting the topic change of knowledge-based systems: a topic-based bibliometric analysis from 1991 to 2016, Knowl. Based Syst. 133 (2017), 255–268.
[24] C. Chen, Z. Hu, S. Liu, H. Tseng, Emerging trends in regenerative medicine: a scientometric analysis in CiteSpace, Expert Opin. Biol. Therapy. 12 (2012), 593–608.
[25] E. Yan, Y. Ding, Q. Zhu, Mapping library and information science in China: a coauthorship network analysis, Scientometrics. 83 (2009), 115–131.
[26] E. Yan, R. Guns, Predicting and recommending collaborations: an author-, institution-, and country-level analysis, J. Informet. 8 (2014), 295–309.
[27] T. Tang, D. Popp, The learning process and technological change in wind power: evidence from China's CDM wind projects, J. Policy Anal. Manag. 35 (2016), 195–222.
[28] Y. Zhang, A. Porter, S.W. Cunningham, D. Chiavetta, N. Newman, Parallel or intersecting lines? Intelligent bibliometrics for investigating the involvement of data science in policy analysis, IEEE Trans. Eng. Manag. (2020), 1–13.
[29] W. Pedrycz, Computational Intelligence: an Introduction, Boca Raton, Florida, USA: CRC Press, 1997. ISBN 9780849326431.
[30] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature. 521 (2015), 436–444.
[31] L.A. Zadeh, Fuzzy sets, Inf. Control. 8 (1965), 338–353.
[32] J. Ma, J. Lu, G. Zhang, Decider: a fuzzy multi-criteria group decision support system, Knowl. Based Syst. 23 (2010), 23–31.
[33] Z. Zhang, H. Lin, K. Liu, D. Wu, G. Zhang, J. Lu, A hybrid fuzzy-based personalized recommender system for telecom products/services, Inf. Sci. 235 (2013), 117–129.

[34] D. Wu, G. Zhang, J. Lu, A fuzzy preference tree-based recommender system for personalized business-to-business e-services, IEEE Trans. Fuzzy Syst. 23 (2014), 29–43.
[35] Y. Ju, A. Wang, X. Liu, Evaluating emergency response capacity by fuzzy AHP and 2-tuple fuzzy linguistic approach, Expert Syst. Appl. 39 (2012), 6972–6981.
[36] T. Back, U. Hammel, H.-P. Schwefel, Evolutionary computation: Comments on the history and current state, IEEE Trans. Evol. Comput. 1 (1997), 3–17.
[37] Y. Zhang, A.L. Porter, Z. Hu, Y. Guo, N.C. Newman, "Term clumping" for technical intelligence: a case study on dye-sensitized solar cells, Technol. Forecast. Soc. Change. 85 (2014), 26–39.
[38] E.C. Noyons, A.F. van Raan, Monitoring scientific developments from a dynamic perspective: self-organized structuring to map neural network research, J. Am. Soc. Inf. Sci. 49 (1998), 68–81.
[39] M. Callon, J.-P. Courtial, W.A. Turner, S. Bauin, From translations to problematic networks: an introduction to co-word analysis, Soc. Sci. Inf. 2 (1983), 191–235.
[40] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Auckland, New Zealand, 1986. ISBN 0-07-054484-0.
[41] L. Waltman, N.J. Van Eck, A smart local moving algorithm for large-scale modularity-based community detection, Eur. Phys. J. B. 86 (2013), 471.
[42] Y. Zhang, L. Shang, L. Huang, A.L. Porter, J. Lu, D. Zhu, A hybrid similarity measure method for patent portfolio analysis, J. Inf. 10 (2016), 1108–1130.
[43] L. Huang, Y. Zhu, Y. Zhang, X. Zhou, X. Jia, A link prediction-based method for identifying potential cooperation partners: a case study on four journals of informetrics, in 2018 Portland International Conference on Management of Engineering and Technology (PICMET), IEEE, Honolulu, HI, USA, 2018, pp. 1–6.
[44] A.F. van Raan, Sleeping beauties in science, Scientometrics. 59 (2004), 467–472.
[45] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in Advances in Neural Information Processing Systems, Lake Tahoe, Nevada, USA, 2013, pp. 3111–3119.
[46] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Trans. Evol. Comput. 6 (2002), 182–197.
[47] N. Srinivas, K. Deb, Muiltiobjective optimization using nondominated sorting in genetic algorithms, Evol. Comput. 2 (1994), 221–248.
[48] J.W. Knowles, E.A. Ashley, Cardiovascular disease: the rise of the genetic risk score, PLoS Med. 15 (2018), e1002546.
[49] S.-W. Hung, A.-P. Wang, Examining the small world phenomenon in the patent citation network: a case study of the radio frequency identification (RFID) network, Scientometrics. 82 (2010), 121–134.
[50] L. Lü, T. Zhou, Link prediction in weighted networks: the role of weak ties, Europhys. Lett. 89 (2010), 18001.